Complexity and word (clump) abundance

To explore the association of low-complexity regions with either the over- or under-representation of words, we examined the ratio of observed to expected words in each of four complexity classes Fig. S1) for Human RefSeq and non-redundant PfamAB proteins. In the evolutionarily redundant Human-RefSeq protein set Fig. S1A, many protein words are over-represented in each of the complexity classes. With non-redundant Pfam-AB (Fig. S1B), the difference in spread between the random MC(1) counts and the higher complexity ($W_2X_1Y_1$,$W_1X_1Y_1Z_1$) counts is greatly reduced, but the differences of the medians remain for the lower complexity (W_4 – W_2X_2) classes. Thus, `seg` has removed many, but not all, of the over-represented low complexity words in low complexity regions.

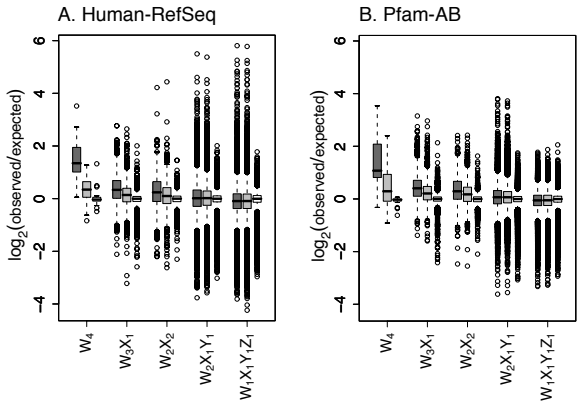

Fig. S1. *Complexity dependence of over-/under-represented words* — Boxplots of the log_2 ratio of observed to estimated expected clump counts are grouped by complexity type for (A) Human-RefSeq and (B) Pfam-AB proteins. The least complex words (W_4) are on the left; the most complex words ($W_1X_1Y_1Z_1$) on the right. Within each complexity type are 3 boxes representing (1) the full library (dark gray, left), (2) the corresponding library after scanning with `seg` to remove low-complexity regions (gray, center), (3) and a random MC(1) library (right). The box-plots display the median as a line within the box, the interquartile distance as the box height, and the whiskers as lines that are 1.5 times the interquartile distance.